Jurnal Formatif 2(2): 140-148
ISSN: 2088-351X

KRITERIA EMPIRIK DALAM MENENTUKAN UKURAN SAMPEL PADA PENGUJIAN HIPOTESIS STATISTIKA DAN ANALISIS BUTIR

IDRUS ALWI
benmashoor@yahoo.com
Program Studi Pendidikan Matematika, Fakultas Teknik, Matematika & Ilmu Pengetahuan Alam - Universitas Indraprasta PGRI Jakarta

Abstrak. Penentuan ukuran sampel sering merupakan langkah penting dalam perencanaan studi statistik dan analisis item dan biasanya sulit. Di antara rintangan penting yang harus dilampaui, seseorang harus memperoleh perkiraan satu atau lebih variasi kesalahan, dan menentukan ukuran efek penting. Ada godaan untuk mengambil jalan pintas. Makalah ini menawarkan beberapa saran untuk sukses dan bermakna dalam menentukan jumlah sampel. Juga dibahas bahwa ukuran sampel bukanlah masalah utama, karena tujuan sebenarnya adalah untuk merancang sebuah studi berkualitas tinggi. Pada akhirnya, kritik membuat beberapa saran buruk berkaitan dengan kekuasaan dan jumlah sampel.

Kata Kunci: jumlah sampel, analisis item, studi statistik.

Abstract. Sample-size determination is often an important step in planning statistical study and item analysis and it is usually a difficult one. Among the important hurdles to surpassed, one must obtain an estimate of one or more error variances, and specify an effects size of importance. There is the temptation to take some shortcuts. This paper offers some suggestions for successful and meaningful sample-size determination. Also discussed is the possibility that sample-size may not be the main issue, that the real goal is to design a high-quality study. Finnaly, criticism is made of some ill-advised shortcuts relating to power and sample-size.

Keyword: sample-size, item analysis, statistical study.

PENDAHULUAN

Salah satu di antara pertanyaan yang sering dikemukakan para peneliti adalah berapa besar jumlah subjek yang perlu ditentukan sebagai sampel. Secara teknis, besarnya sampel tergantung pada ketepatan yang diinginkan peneliti dalam menduga parameter populasi pada taraf kepercayaan tertentu. Tidak ada satu kaidah pun yang dapat digunakan untuk menetapkan besarnya sampel. Akan tetapi secara empirik perkiraan besarnya sampel yang dibutuhkan dapat ditentukan. Berapa ketentuan sampel yang dibutuhkan secara empirik banyak dibahas oleh para peneliti. Kesepakatan para peneliti dalam menentukan besarnya sampel sebagai kriteria empirik dalam menguji hipotesis statistika akan sangat bermanfaat bagi peneliti saat ini sebagai pedoman dalam menentukan ukuran sampel dalam penelitian.

PEMBAHASAN

Persyaratan Ukuran Sampel Dalam Pengujian Hipotesis Statistika.

Sebelum memilih subjek yang akan dilibatkan dalam penelitiannya, seorang peneliti harus terlebih dahulu menentukan jumlahnya. Penentuan berapa besar jumlah

subjek yang diperlukan, seringkali menjadi permasalahan dalam merencanakan suatu penelitian. Tidak ada aturan pasti berapa banyak agar sampel dapat mewakili populasi. Akan tetapi, secara umum dapat dikatakan bahwa semakin besar sampel semakin besar kemungkinan dapat mencerminkan populasi.

Secara statistika dinyatakan bahwa ukuran sampel yang semakin besar diharapkan akan memberikan hasil yang semakin baik. Dengan sampel yang besar, mean dan standar deviasi yang diperoleh mempunyai probabilitas yang tinggi untuk menyerupai mean dan standar deviasi populasi. Hal ini karena jumlah sampel ada kaitannya dengan pengujian hipotesis statistika. Meskipun sampel yang besar akan semakin baik, sampel yang kecil bila dipilih secara acak dapat mencerminkan pula populasi dengan akurat (Hajar, 1996: 147).

Membahas masalah ukuran sampel, maka dapat dikemukakan suatu teorema tentang variabel tunggal atau univariat, yaitu teorema limit sentral, yang menyatakan statistik rata-rata mempunyai distribusi normal untuk ukuran sampel yang mendekati tak berhingga. Akan tetapi dalam praktek, teorema limit sentral telah dapat diterapkan untuk ukuran sampel minimal 30. Bahkan dinyatakan untuk ukuran sampel lebih besar dari 20, distribusi normal telah dapat dipakai untuk mendekati distribusi binomial (Agung, 2006: 83). Ukuran sampel lebih besar daripada 30 dan lebih kecil daripada 500, cocok dipakai untuk kebanyakan penelitian. Jika sampel harus dibagi dalam dua kategori seperti laki-laki dan perempuan, maka diperlukan ukuran sampel minimal 30 untuk setiap kategori (Agung, 2005: 113)

Menurut Gay, Mills dan Airasian (2009: 133) untuk penelitian metode deskriptif, minimal 10% populasi, untuk populasi yang relatif kecil minimal 20%, sedangkan untuk penelitian korelasi diperlukan sampel sebesar 30 responden. Untuk penelitian eksperimen dan komparatif diperlukan sampel 30 responden untuk setiap kelompok yang akan dibandingkan. Khusus untuk penelitian eksperimen dan komparatif, menurut Borg and Gall (2007: 176) diperlukan sampel 15-30 responden setiap kelompok. Menurut Krejcie dan Morgan dalam Schreiber dan Asner-Self (2011: 92) untuk populasi di bawah 100 diambil semua, bila populasi berjumlah 500 diambil 50%, bila populasi berjumlah 5000 diambil 357 responden, bila populasi berjumlah 100.000 diambil 384 responden.

Menurut Agung, secara empiris statistik rata-rata mempunyai distribusi normal dengan memakai 1000 buah sampel yang dipilih secara random. Akan tetapi, untuk data atau variabel multivariat belum terdapat kesepakatan dan masih merupakan ketentuan yang sangat subjektif. Dalam penelitian multivariat, maka ukuran sampel harus beberapa kali lebih besar daripada jumlah variable penelitian yang ditinjau. Untuk eksperimen sederhana dengan kendali ketat keberhasilan penelitian dapat dicapai dengan memakai sampel berukuran 10 sampai dengan 20.

Dalam penelitian pendidikan, terutama dalam penelitian eksperimen, probabilitas sampling tidak selalu diperlukan atau mungkin tidak dapat dilakukan pemilihan subjek dari populasi yang lebih besar. Dalam hal yang demikian, peneliti biasanya menggunakan sampling tersedia (*availability sampling*), yakni peneliti memanfaatkan subjek yang tersedia, misalnya sekelompok siswa dalam kelas tertentu. (Hajar, 1996: 147)

Penggunaan jumlah sampel yang berbeda dari populasi yang sama tidak menghasilkan perbedaan yang berarti. Hasil dari sampel yang hanya dua persen tidak jauh berbeda dengan hasil yang menggunakan sampel sepuluh persen dari populasi. Pada penelitian eksperimen yang dikontrol secara ketat, bila masing-masing kelompok terdiri antara 8 sampai 10 subjek sudah dianggap memadai untuk mendapatkan hasil yang akurat, walaupun pengujian statistik selalu memperlihatkan signifikansi apabila ukuran sampel cukup besar (Holland & Wainer, 1993: 12).

Banyak penelitian eksperimen menggunakan ukuran sampel minimal yang sangat kecil, yaitu 3-5 unit observasi dalam setiap sel atau kelompok yang diperhatikan. Akan tetapi untuk kemudahan menghitung nilai rata-rata dan standar deviasinya disarankan untuk menggunakan ukuran minimal 5.

Berdasarkan keterangan di atas, dengan sendirinya kita tidaklah mungkin menentukan suatu ukuran sampel tertentu yang dapat dinyatakan paling cocok atau terbaik untuk suatu penelitian. Walaupun terdapat rumus-rumus yang dapat dipakai untuk memperkirakan ukuran sampel yang diperlukan, akan tetapi ukuran sampel yang diperoleh tersebut hanya merupakan suatu pedoman, bukanlah merupakan syarat yang absolut.

Ukuran sampel Dalam Ujicoba Instrumen.

Instrumen penelitian memegang peranan penting dalam penelitian kuantitatif karena kualitas data yang diperoleh dalam banyak hal ditentukan oleh kualitas instrumen yang dipergunakan. Jika instrumen yang digunakan dapat dipertanggungjawabkan, data yang diperoleh juga dapat dipertanggungjawabkan. Artinya, data yang bersangkutan dapat mewakili dan atau mencerminkan keadaan sesuatu yang diukur pada diri subjek penelitian.

Untuk mengetahui baik tidaknya instrumen penelitian dilakukan melalui analisis tes. Hasil analisis tes dapat digunakan untuk menguji apakah instrumen berfungsi dengan baik. Di samping itu, hasil analisis tes dapat digunakan untuk mengetahui apakah butir tes termasuk kategori baik, perlu diperbaiki, atau jelek. Baik tidaknya suatu instrumen dapat dianalisis melalui indikator-indikator yang merupakan unsur-unsur dari kualitas tes, yaitu: (1) validitas/kesahihan, (2) reliabilitas/keandalan, (3) daya pembeda, (4) tingkat kesukaran, dan (5) efektivitas pengecoh.

Validitas

Pengujian validitas berkaitan dengan permasalahan apakah instrumen yang dimaksud untuk mengukur sesuatu itu, memang dapat mengukur secara tepat sesuatu yang akan diukur tersebut. Pengujian validitas atau yang dikenal dengan telaah mutu soal dilakukan sebelum soal diujikan kepada pihak yang dijadikan subjek penelitian.

Dari cara estimasinya yang disesuaikan dengan sifat dan fungsi setiap tes, tipe validitas pada umumnya digolongkan dalam tiga kategori, yaitu *content validity* (validitas isi), *construct validity* (validitas konstruk) dan *criterion-related validity* (validitas berdasarkan kriteria). *Content validity* atau validitas isi dilakukan melalui telaah kualitatif, sedangkan *construct validity* (validitas konstruk) dan *criterion-related validity* (validitas berdasarkan kriteria) dapat dilakukan melalui telaah kuantitatif atau teknik analisis statistika.

Telaah Kualitatif

Secara kualitatif, analisis soal dilakukan berdasarkan pertimbangan (*professional judgment*) ahli materi, ahli konstruksi dan ahli bahasa. Ada beberapa teknik yang dapat digunakan untuk telaah mutu soal secara kualitatif, diantaranya adalah teknik moderator dan teknik panel.

Teknik moderator merupakan teknik berdiskusi yang di dalamnya terdapat satu orang sebagai penengah. Berdasarkan teknik ini, setiap butir soal didiskusikan secara bersama-sama dengan beberapa ahli seperti guru yang mengajarkan materi, ahli materi, penyusun/pengembang kurikulum, ahli penilaian, ahli bahasa, berlatar belakang psikologi. Teknik ini sangat baik karena setiap butir soal dilihat secara bersama-sama berdasarkan kaidah penulisannya. Di samping itu, para penelaah dipersilahkan

mengomentari/ memperbaiki berdasarkan ilmu yang dimilikinya. Setiap komentar/masukan dari peserta diskusi dicatat oleh notulis. Setiap butir soal dapat dituntaskan secara bersama-sama, perbaikannya seperti apa. Namun, kelemahan teknik ini adalah memerlukan waktu lama untuk mendiskusikan setiap satu butir soal.

Teknik panel merupakan suatu teknik menelaah butir soal yang setiap butir soalnya ditelaah berdasarkan kaidah penulisan butir soal, yaitu ditelaah dari segi materi, konstruksi, bahasa/budaya, kebenaran kunci jawaban/pedoman penskorannya yang dilakukan oleh beberapa penelaah. Caranya adalah beberapa penelaah diberikan: butir-butir soal yang akan ditelaah, format penelaahan, dan pedoman penilaian/ penelaahannya. Pada tahap awal para penelaah diberikan pengarahan, kemudian tahap berikutnya para penelaah berkerja sendiri-sendiri di tempat yang tidak sama. Para penelaah dipersilakan memperbaiki langsung pada teks soal dan memberikan komentarnya serta memberikan nilai pada setiap butir soalnya yang kriterianya adalah: baik, diperbaiki, atau diganti.

Secara ideal penelaah butir soal di samping memiliki latar belakang materi yang diujikan, beberapa penelaah yang diminta untuk menelaah butir soal memiliki keterampilan, seperti guru yang mengajarkan materi itu, ahli materi, ahli pengembang kurikulum, ahli penilaian, psikolog, ahli bahasa, ahli kebijakan pendidikan, atau lainnya.

Telaah Kuantitatif

Secara kuantitatif atau empirik, pengujian butir instrumen atau soal tes dapat dilakukan dengan teknik analisis statistika. Misalnya untuk uji validitas dengan cara menghitung koefisien korelasi antara sekor butir instrumen atau soal tes dengan sekor total instrumen atau tes. Butir atau soal yang dianggap valid adalah butir instrumen atau soal tes yang sekornya mempunyai koefisien korelasi yang signifikan dengan skor total instrumen atau tes. Rumus statistika yang banyak digunakan diantaranya korelasi *Point Biserial* untuk data dikotomi dan korelasi *Product Moment* untuk data kontinum.

Dalam telaah kuantitatif, jumlah butir dan ukuran responden berpengaruh terhadap hasil analisis yang dilakukan. Untuk ketelitian dan keakuratan hasil, diperlukan minimal besarnya jumlah butir dan ukuran responden. Mengantisipasi banyaknya butir yang akan gugur, disarankan untuk melipatgandakan jumlah butir yang akan digunakan sebagai instrumen penelitian. Jika jumlah yang akan digunakan dalam penelitian sebanyak 20 butir, maka butir yang diujicobakan dapat berjumlah 40 atau dua kali lipat.

Dalam hal jumlah responden, Crocker dan Algina (1986: 322) menyatakan bahwa demi kestabilan, minimal diperlukan 200 responden. Nunnaly (1970: 214-215) menyatakan bahwa ukuran responden pada ujicoba adalah sebesar sepuluh kali dari jumlah butir di dalam alat ukur. Alat ukur dengan 30 butir misalnya, memerlukan 10 x 30 = 300 responden. Banyak alat ukur yang terdiri atas 30 sampai 60 butir sehingga di dalam ujicobanya, alat ukur ini memerlukan 200 sampai 600 responden. Tetapi dalam kondisi tertentu, untuk 20 butir soal dapat digunakan 100 responden.

Untuk menentukan valid atau tidak valid suatu butir soal, maka diperlukan interprestasi koefisien validitas. Interpretasi koefisien validitas bersifat relatif, artinya tidak ada batasan pasti mengenai koefisien terendah yang harus dipenuhi agar validitas dinyatakan memuaskan.

Koefisien validitas yang baik, setinggi mungkin mendekati harga r_{xy} = 1,00. Akan tetapi untuk memperoleh koefisien validitas yang tinggi lebih sulit daripada memperoleh koefisien reliabilitas yang tinggi. Hal ini menjadikan alasan setiap penulis butir soal untuk bersikap realistik dan tidak menuntut koefisien yang setinggi koefisien reliabilitas.

Di bawah ini adalah tabel interpretasi koefisien validitas yang diambil dari beberapa tes yang telah dilakukan, sebagai berikut:

Tabel 1. Interpretasi Koefisien Validitas

Nama Tes	Batas Valid/ Tidak Valid
American College Testing Program Composite	0,27
College Entrance Examination Board SAT	0,29
Differentian Aptitude Test: Verbal Reasoning	0,39
Percentill Rank in High School Graduation Class	0,43

Suatu kesepakatan umum menyatakan bahwa koefisien validitas dapat dianggap memuaskan apabila melebihi r_{xy} = 0,30. Siapapun boleh menerima atau menolak batasan ini karena memang penetapan angka tersebut tidak didasari logika matematika melainkan merupakan konvensi tidak tertulis yang didasari oleh pertimbangan professional dan pengalaman saja.

1) Uji Validitas Untuk Butir Soal Bentuk Pilihan Ganda menggunakan rumus ***Point Biserial***

$$r_{pbi} = \frac{\bar{X}_p - \bar{X}_q}{s}\sqrt{pq} \text{ atau } r_{pbi} = \frac{\bar{X}_p - \bar{X}_t}{s}\sqrt{\frac{p}{q}}$$

$\bar{X}p$: rata-rata skor kemampuan peserta didik yang menjawab benar
$\bar{X}q$: rata-rata skor kemampuan peserta didik yang menjawab salah
$\bar{X}t$: rata-rata skor dari skor total
S : simpangan baku skor total
p : proporsi jawaban benar terhadap semua jawaban siswa
q : 1-p

2) Uji Validitas Untuk Butir Soal Skala Kontinum (Uraian dan Non-Tes):
Sekor butir instrumen atau soal tes kontinum (misalnya bentuk soal Uraian dan skala sikap dengan sekor butir 0 – 10 atau 1- 5) dan diberi symbol x_i dan sekor total instrument atau tes diberi symbol x_t, maka rumus yang digunakan untuk menghitung koefisien korelasi antara sekor butir instrumen atau soal dengan sekor total instrumen atau sekor total tes adalah rumus ***Product Moment***:

Rumus: $$r_{xy} = \frac{N\sum X_1 X_2 - (\sum X_1)(\sum X_2)}{\sqrt{\{N\sum {X_1}^2 - (\sum X_1)^2\}\{N\sum {X_2}^2 - (\sum X_2)^2\}}}$$

Di mana r_{xy} adalah koefisien korelasi yang dicari, N adalah *Number of Cases*, X_1 adalah sekor butir dan X_2 adalah sekor total.

Reliabilitas

Reliabilitas adalah sejauhmana hasil suatu pengukuran dapat dipercaya. Hasil pengukuran dapat dipercaya apabila dalam beberapa kali pelaksanaan pengukuran terhadap kelompok subjek yang sama diperoleh hasil yang relatif sama, selama aspek yang diukur dalam diri subjek memang belum berubah. Ada beberapa prosedur untuk menghitung indeks reliabilitas tes, di antaranya melalui pendekatan tes ulang (test-retest), pendekatan bentuk paralel, dan pendekatan konsistensi internal.

Di antara pendekatan konsistensi internal adalah metode Kuder-Richardson 20 (KR-20) dan *Alpha Cronbach*. Menurut Nitko (1983: 395) Kuder-Richardson 20 (KR-20) digunakan untuk menghitung nilai reliabilitas tes dalam bentuk tes objektif yang hanya

menggunakan sekor dikotomi, yaitu bila benar = 1 dan salah = 0, seperti pada bentuk tes pilihan ganda. Sedangkan koefisien *Alpha Cronbach* digunakan untuk menghitung nilai reliabilitas tes dalam bentuk uraian atau skala sehingga pengukurannya tidak hanya menggunakan sekor benar = 1 dan salah = 0, seperti pada tes objektif, melainkan dapat menggunakan sekor 1 – 10 atau skala 1 – 5, dan sebagainya. Adapun rumus:

Kuder-Richardson 20 (KR-20): $r = \frac{k}{k-1}\left(1 - \frac{\sum pq}{s^2}\right)$ dan,

Koefisien *Alpha Cronbach* : $r = \frac{k}{k-1}\left(1 - \frac{\sum S_b^2}{S_t^2}\right)$.

Selanjutnya, untuk menentukan reliabel atau tidak reliabel suatu tes, maka diperlukan interprestasi koefisien reliabilitas. Sebagaimana pada interpretasi validitas, interpretasi terhadap koefisien reliabilitas juga bersifat relatif, tidak ada batasan pasti mengenai koefisien terendah yang harus dipenuhi agar suatu pengukuran dapat disebut reliabel. Terdapat dua kriteria empirik untuk menentukan besarnya koefisien reliabilitas yang memadai. Kriteria empirik pertama berkenaan dengan bidang ilmu, dan kriteria empirik kedua berkenaan dengan statistika.

Pada umumnya, untuk bidang ilmu yang memiliki pengukuran dengan kecermatan tinggi seperti pengukuran keberhasilan belajar matematika yang baku memiliki koefisien reliabilitas yang tinggi yakni di atas 0,90. Dengan demikian, koefisien reliabilitas yang memadai pada ujian keberhasilan matematika adalah sekitar 0,90. Menurut Ebel (1979: 275) suatu koefisien reliabilitas di sekitar 0,90 atau lebih, dapat dianggap memuaskan.

Sebaliknya untuk bidang ilmu yang belum memiliki kecermatan pengukuran yang tinggi, koefisien reliabilitas yang rendah pun sudah dianggap memadai. Hal ini dapat diperiksa pada jurnal ilmu bersangkutan. Jika di dalam jurnal bidang ilmu itu ditemukan bahwa koefisien reliabilitas pada pengukurannya di sekitar 0,40 maka koefisien reliabilitas yang memadai adalah 0,40 (Naga, 2009: 93).

Secara statistika, koefisien reliabilitas yang memadai adalah koefisien korelasi linear yang memadai. Kriteria empirik menyatakan bahwa irisan variansi X dan variansi Y yang disebut koefisien determinasi (d) dianggap memadai apabila telah mencapai = 0,50. Koefisien determinasi berkaitan dengan koefisien korelasi linear (ρ_{xy}), maka hubungan keduanya adalah $\rho_{xy} = \sqrt{d}$. Jika d = 0,50 dianggap memadai maka koefisien korelasi linear dengan nilai $\sqrt{d} = \sqrt{0{,}50}$ = 0,71 dianggap sudah memadai. Karena koefisien reliabilitas merupakan jenis koefisien korelasi linear, maka secara statistika koefisien reliabilitas yang memadai adalah **0,71** atau lebih.

Tingkat Kesukaran

Tingkat kesukaran soal adalah peluang untuk menjawab benar suatu soal pada tingkat kemampuan tertentu yang biasanya dinyatakan dalam bentuk indeks. Untuk menghitung tingkat kesukaran soal uraian berbeda dengan cara yang digunakan pada tes objektif. Untuk menghitung tingkat kesukaran ada beberapa cara yaitu: (1) *proportion correct*, (2) indeks kesukaran linier, (3) indeks Davis, dan (4) skala Bivariat.

Untuk bentuk soal Pilihan Ganda, cara yang paling mudah dan paling umum digunakan adalah dengan skala rata-rata atau proporsi menjawab benar atau *proportion correct* (*p*), yaitu jumlah peserta tes yang menjawab benar pada soal yang dianalisis dibandingkan dengan peserta tes seluruhnya. Persamaan yang digunakan untuk menentukan tingkat kesukaran (*p*) ini adalah:

$$\boldsymbol{P = \frac{\sum B}{N}}$$

P : proporsi menjawab benar pada butir tertentu
$\sum B$: banyaknya peserta tes menjawab benar
N : jumlah peserta tes yang menjawab benar

Sedangkan untuk bentuk Soal Uraian, Untuk mengetahui tingkat kesukaran soal bentuk uraian digunakan rumus berikut ini:

$$Tingkat\,Kesukaran = \frac{Mean}{Skor\,maksimum}$$

Besarnya tingkat kesukaran antara 0 dan 1. Tingkat kesukaran dikategorikan menjadi tiga bagian seperti tabel di bawah ini:

Tabel 2. Kategori Tingkat Kesukaran

Proportion Correct (*p*)	Kategori Soal
0,71 - 1,00	Mudah
0,31 - 0,70	Sedang
0,00 - 0,30	Sukar

Daya Pembeda Soal

Daya pembeda atau daya beda suatu butir soal berfungsi untuk menentukan dapat tidaknya suatu butir soal membedakan kelompok dalam aspek yang diukur sesuai dengan perbedaan yang ada pada kelompok itu. Tujuan dari pengujian daya pembeda adalah untuk melihat kemampuan butir soal dalam membedakan antara peserta didik yang berkemampuan tinggi (M_T) dengan peserta didik yang berkemampuan rendah (M_R).

Kelley dalam Naga (2009: 89) menemukan bahwa nilai optimal penggunaan ukuran kelompok adalah $M_T = M_R = 27\%$. Sejak itulah banyak orang menentukan pilahan 27% untuk kelompok tinggi dan 27% untuk kelompok rendah pada ukuran responden yang besar. Selanjutnya pemilahan responden ke kelompok tinggi dan kelompok rendah dilakukan melalui:

- untuk responden (M) < 371, $M_T=M_R=50\%$
- untuk responden (M) ≥ 371, $M_T=M_R=27\%$

Daya diskriminasi yang baik memang pada umumnya terdapat pada item yang tidak terlalu mudah dan juga tidak terlalu sukar, yaitu apabila harga p berkisar antara 0,40 sampai dengan 0,60. Dalam seleksi item, setiap item yang memiliki indeks diskriminasi lebih besar dari 0,50 dapat langsung dianggap baik, item yang memiliki indeks diskriminasi kurang dari 0,20 dapat langsung dibuang, sedangkan item lainnya dapat ditelaah lebih lanjut untuk direvisi.

Secara empirik, minimum tingkat daya pembeda yang memadai seperti tabel di bawah ini:

Tabel 3. Kategori Daya Beda

Nama Ahli	Daya Beda Minimum
Crocker & Algina (1986: 324)	0,2
Nunnaly (1970: 202)	0,2
Aiken (1994: 65)	0,2
Mehrens & Lehmans (1991: 167)	0,2

Kesepakatan beberapa ahli di atas menyatakan bahwa koefisien indek diskriminasi dapat dianggap baik apabila melebihi 0,20.

Efektifitas Pengecoh

Efektifitas pengecoh merupakan salah satu faktor yang dijadikan dasar dalam penelaahan soal. Hal ini dimaksudkan untuk mengetahui berfungsi tidaknya pilihan jawaban yang tersedia selain kunci jawaban. Suatu pilihan jawaban (pengecoh) dapat dikatakan berfungsi apabila pengecoh paling tidak dipilih oleh 5 % peserta tes/siswa.

Selain cara klasik seperti yang telah diuraikan di atas, penganalisisan tes dan penggunaan data hasil analisis dapat dilakukan dengan pendekatan teori tes modern atau *Item Response Theory*. Dalam konsep *Item Response Theory* (IRT) setiap soal diwakili oleh *Item Characteristic Curve*. Pada Item Response Theory, parameter soal yang dihitung tergantung kepada model parameter yang dikehendaki. Model satu parameter atau biasa disebut *Model Rasch* hanya mempunyai satu parameter yaitu indeks kesukaran soal. Dua parameter model, yaitu tingkat kesukaran dan daya pembeda soal, sedangkan tiga parameter model terdiri dari tingkat kesukaran, daya pembeda dan *pseudo guessing*.

Konsep *Item Response Theory* (IRT) sangat berguna untuk memecahkan masalah-masalah dalam penyeleksian soa-soal dalam rangka mendisain suatu perangkat tes tertentu. Salah satu keunggulan utama *Item Response Theory* (IRT) dibandingkan teori Tes Klasik adalah dalam konsep *Item Response Theory* statistik soal seperti tingkat kesukaran, daya pembeda terletak dalam skala yang sama dengan kemampuan siswa yang diukur.

Item Response Theory dikembangkan melihat adanya beberapa kelemahan yang terdapat pada teori tes klasik. Dalam teori tes klasik, tingkat kesukaran soal biasanya dilaporkan dalam skala 0 sampai 1 dan didefinisikan atau dihubungkan dengan populasi pengikut tes atau proporsi banyaknya pengikut tes yang menjawab soal tertentu dengan benar. Sedangkan domain kemampuan siswa, walaupun juga dilaporkan dalam skala 0 sampai 1 tetapi didefinisikan atau dihubungkan dengan populasi soal dalam tes. Jadi terlihat bahwa yang menjadi masalah adalah skala tingkat kesukaran soal tidak sama atau berbeda definisinya dengan skala domain kemampuan siswa. Dengan kata lain, kalau kita berbicara mengenai tingkat kesukaran soal maka populasinya adalah kumpulan orang-orang pengikut tes, tetapi kalau kita berbicara mengenai domain kemampuan siswa, maka populasinya adalah kumpulan materi yang dites.

Perbedaan yang utama dari kedua model teori tes tersebut adalah terletak pada kelompok sampel. Tes yang telah disusun hanya dapat diterapkan secara baik kepada individu-individu yang kondisinya sama dengan kondisi kelompok yang dijadikan subjek dalam pengembangan tes. Seperti yang dikatakan Azwar, bahwa parameter-parameter item dalam teori klasik merupakan karakter item yang tergantung pada kelompok sampel yang digunakan untuk menghitungnya.

Dalam penggunaan alat ukur yang termasuk dalam *Item Response Theory*, Crocker dan Algina (1986: 322) merekomendasikan minimal 200 responden. Wright dan Stone merekomendasikan minimal panjang tes 20 butir dan 200 responden. Hullin, Lissak dan Drasgow (1983: 99) merekomendasikan panjang tes 30 butir dan 500 responden untuk dua parameter model (L2P), 60 butir dan 1000 sampel untuk tiga parameter butir (L3P).

Berbeda dengan pendekatan IRT, untuk mendapatkan keakuratan alat ukur dalam mendeteksi keberbedaan fungsi butir (DIF), model klasik memiliki keuntungan tidak mempersyaratkan ukuran sampel yang besar. Bagi model klasik makin besar sampel makin baik. Berbeda dengan IRT, jika ukuran sample pada kelompok fokus kecil, hal ini merupakan problem. Keakuratan penggunaan alat ukur untuk mendeteksi keberbedaan

fungsi butir pada IRT , tergantung model logistik parameter yang akan digunakan. Untuk tiga parameter logistik, maka DIF akan akurat jika menggunakan 1000 responden dan 60 butir (Wiberg, 2007).

PENUTUP

Dari uraian di atas dapat disimpulkan bahwa ukuran sampel menjadi salah satu hal penting dalam pelaksanaan penelitian, baik dalam melakukan uji hipotesis maupun dalam melakukan analisis butir. Pada prinsipnya, semakin banyak sampel semakin baik hasil penelitian.

DAFTAR PUSTAKA

Agung, I Gusti Ngurah. 2006. **Statistika Penerapan Model Rerata Sel Multivariat dan Model Ekonometri dengan SPSS**. Jakarta: Yayasan SAD Satria Bhakti.

Aiken, Lewis R. 1997. **Psychological Testing and Assesment**. Boston: Allyn and Bacon.

Azwar, Saifuddin. 1997. **Reliabilitas dan Validitas**. Yogyakarta: Pustaka Pelajar.

Borg, Walter R, Meredith D, Gall and Joyce P. Gall. 2007. **Education Research**. New York: Pearson Education, Inc.

Crocker, Linda dan James Algina. 1986. **Introduction To Classical And Modern Test Theory.** New York: Holt: Rinehart and Winston.

Ebel, Robert L. 1979. **Essential of Educational Measurement**. New Jersey: Prentice-Hall. Inc.

Gay, LR, Geoffrey E. Mills and Peter Airasian. 2009. **Educational Research, Competencies for Analysis and Application**. New Jersey: Pearson Education, Inc.

Hajar, Ibnu. 1996. **Dasar-Dasar Metodologi Penelitian Kwantitatif Dalam Pendidikan**. Jakarta: Raja Grafindo Persada.

Holland, Paul W & Howard Wainer (ed). 1993. **Differential Item Functioning**. New Jersey: Lawrence Erlbaum Associates Publisher.

Hulin, Charles L, Pritsz Drasgow and Charles K Parsons. 1983. **Item Respon Theory, Aplications to Psychological Measurement**. Illinois: Dow Jones-Irwin.

McMillan, James H. 2008. **Education Research, Fundamentals For The Consumer**. New York: Pearson Education, Inc.

Mehrens, William A and Irvin J. Lehmans. 1991. **Measurement and Evaluation in Education and Psychology.** Fort Worth: Harcourt Brace College Publisher.

Naga, Dali S. 2009. **64 Rumus Terapan Probabilitas dan Sekor Pada Hipotesis Statistika**. Jakarta: Grasindo.

Nitko, Anthony J. 1983. **Educational Test And Measurement An Introduction**. New York: Harcourt Brace Jovanovich.

Nunnaly, Jum C, Jr. 1970. **Introduction to Psychological Measurement**. New York: McGraw-Hill Book Company.

Schreiber, James and Kimberly Asner-Self. 2011. **Educational Research**. New Jersey: John Wiley & Sons, Inc.